Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Guangyi Zhang

Personalized Treatment-Response Trajectories:

Errors-in-variables, Interpretability, and Causality

Master's Thesis
Espoo, March 27, 2019

DRAFT! — March 27, 2019 — DRAFT!

Supervisors: Professor Pekka Marttinen, Aalto University
Advisor: Professor Pekka Marttinen

Aalto University
School of Science

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

Author:	Guangyi Zhang		
Title:	Personalized Treatment-Response Trajectories: Errors-in-variables, Interpretability, and Causality		
Date:	March 27, 2019	**Pages:**	40
Major:	Computer Science	**Code:**	SCI3042
Supervisors:	Professor Pekka Marttinen		
Advisor:	Professor Pekka Marttinen		

One fundamental problem in many applications is to estimate treatment-response trajectories given multidimensional treatment variables. However, in reality, the estimation suffers severely from measurement error both in treatment timing and covariates, for example when the treatment data are self-reported by users. We introduce a novel data-driven method to tackle this challenging problem, which models personalized treatment-response trajectories as a sum of a parametric response function, based on restored true treatment timing and covariates and sharing information across individuals under a hierarchical structure, and a counterfactual trend fitted by a sparse Gaussian Process. In a real-life dataset where the impact of diet on continuous blood glucose is estimated, our model achieves a superior performance in estimation accuracy and prediction.

Keywords:	bayesian methods, errors-in-variables, causality, hierarchical modeling
Language:	English

Acknowledgements

I wish to thank my supervisor Professor Pekka Marttinen, without whom this thesis would have not been successful. I am extremely fortunate to have worked with him for over a year, during which he constantly encouraged me when in difficulty, inspired me when the work went hopeless, and kindly criticized my work when I felt self-satisfied. He not only taught me how to do research, but also set up an upright and kind role model for me. I can by no means express too much gratitude towards him. I would also like to thank postdoctoral researcher Reza Ashrafi for his great efforts in perfecting this work, and Aalto CS-IT and Science-IT for providing powerful and helpful computational resources.

Personally, I would also extend my thanks to all my family, friends and colleagues who offer me their generous support during my study. Particularly, I am very fortunate to study with Zheyang in my master, whose genius and independence spur me on to keep pursuing intellectual enjoyment. Together with our mutual dear friends, Yuanhao and Qianyun, we shielded ourselves from the long cold dark winters in Finland with moments of sheer joy.

Espoo, March 27, 2019

Guangyi Zhang

Abbreviations and Acronyms

GP	Gaussian Process
EIV	Errors-in-variables
RCT	Randomized controlled trials
ATE	Average treatment effect
ITE	Individual treatment effect
NUC	No unmeasured confounders
IP	Inverse probability weighting
RNN	Recurrent neural network
FE	Fixed-effects models
IV	Instrumental variables
MLE	Maximum likelihood estimation
MCMC	Markov Chain Monte Carlo
SE	Squared Exponential
NUTS	No-U-Turn

Contents

Chapter 1

Introduction

Recently, with the increasing amount of *electronic health records* (EHRs), it becomes possible to leverage statistical machine learning techniques to establish more efficient healthcare systems. Examples of applications of machine learning for healthcare include medical imaging, subtypes clustering, and treatment recommendation [Ghassemi et al., 2018], which immensely alleviate pressure on limited medical resources.

In particular, one fundamental problem in healthcare is the estimation of treatment effect, for example, how much dialysis lowers a patient's creatinine. This information is essential for practitioners to prescribe patients an effective treatment in a safe dose. In the past, this was mostly obtained by *randomized control trials* (RCTs), which are costly and often infeasible due to reasons such as medical ethics. While data-driven techniques on time-fixed treatments have been extensively studied, time-varying treatments still poses many unsolved challenges, for example, self-reported noisy treatment timing in data. Treatment effect with a long duration, known as treatment-response trajectories, takes a form of continuous functions instead of scalar values, which requires more sophisticated parametric or nonparametric functional modeling. Personalized estimation also encounters a long-standing problem of data sparseness.

One ambitious goal of machine learning community is to economically estimate treatment-response trajectories by a data-driven approach. Major technical challenges presented in the way include trustworthy machine learning, error-tolerant or robust estimation, and personalization. This thesis proposes a novel model for the estimation of treatment-response trajectories, together with solutions to all above obstacles. The performance of the model is verified on a real-life glucose dataset where patients' diet is considered as a treatment.

1.1 Problem statement

Causality is crucial to trustworthy treatment effect estimation. Estimating treatment effect by a means of curving fitting, a dominant methodology in the realm of machine learning, is straightforward, but may cause fatal mistakes in vital disease diagnosis, because association between the treatment and some unknown factors can result in a deceptive conclusion. Therefore, in order to employ reliable machine learning techniques in healthcare, we have to enrich our mathematical tools to define and explain causality [Pearl, 2009] [Miguel A. Hernán, 2018], and replace traditional statistical associations with them.

Observed data in healthcare are typically with poor quality, such as censoring, missingness, scarcity, heterogeneity, and noise. In such a dreadful circumstance, it is likely to give rise to a biased and meaningless result. While most machine learning models are designed taking into account the error in dependent variables (outcomes), error in independent variables (predictors) has caught little attention. However, the former only induces additional variability in estimation in either linear or nonlinear models [Carroll et al., 2006, Chapter 15] whereas the latter often spawn disastrous consequences–a biased estimation. Thus, *errors-in-variables* (EIV) modeling is an indispensable component in robust estimation.

Last but not least, there exists an enormous variation in treatment effect among different people. Estimating only an average effect across all patients is inadequate, whereas a separate estimation for each individual is unrealistic due to data sparseness. How to efficiently share information across individuals becomes an important topic in personalized healthcare. In the Bayesian paradigm, an elegant solution to this problem is Bayesian hierarchical modeling, assuming individual parameters follow one higher level common distribution.

1.2 Structure of the thesis

Prior work and related background are first reviewed in Chapter 2. Then our method and accompanying experiments are presented in details in Chapter 3 and Chapter 4 respectively. A discussion follows in Chapter 5, after which this thesis concludes with Chapter 6.

Chapter 2

Background

In this section, an introduction to technical background for key elements in our model is presented, including Causality, errors-in-variables modeling, and Gaussian Processes.

2.1 Causality

Causality provides guidance on estimating unbiased treatment effect from observational data, where potential factors that may affect both the treatment and outcome have not been controlled between experimental and control groups. An estimation is likely to be biased without using causal techniques, when the treatment is confounded by a factor that correlates with both treatment and outcome. For example, in estimating the effect of smoking on catching lung cancer, it may be that people who carry a gene that causes lung cancer like smoking more than people who do not.

This causal relation is best demonstrated by a causal graph, which is similar to a graphical model [Bishop and Mitchell, 2014] but with notation carrying different meanings. In a graphical model, one directed edge represents conditional dependence between random variables, while in a causal graph, it means a causal relation. For the example above, its causal graph can be depicted as Figure 2.1a, where A stands for treatment, Y an outcome, Z a observed confounder, and U a unobserved confounder. In the case of an intervention, the causal graph needs to be modified to accommodate such an external force–all arrows pointed to the treatment need deleting, shown in Figure 2.1b. That is to say, the $E[Y|A = a]$ estimated based on the modified causal graph represents a causal outcome Y given treatment $A = a$, equivalent to the one estimated using RCTs, which is termed a potential outcome and is denoted as $E[Y^{A=a}]$. With potential outcomes given differ-

ent treatment, *average treatment effect* (ATE) is typically used to determine the effectiveness of a treatment, which is defined as follows with a binary treatment.

$$ATE = E[Y^{A=1}] - E[Y^{A=0}] \tag{2.1}$$

Fortunately, $E[Y^{A=a}]$ in the modified causal graph can be estimated using components from the original causal graph with proper assumptions. An unbiased estimation of ATE can be obtained as long as all confounders are measured, known as *No Unmeasured Confounders* (NUC) assumption. For the example shown in Figure 2.1a, this means U needs to be measured in order to make the causal estimation possible, thus turned into Figure 2.1c.

$$P(Y^{A=a} = y) = \sum_z P(Y = y|A = a, Z = z)P(Z = z) \tag{2.2}$$

where every term happens in the counterfactual world, but $P(Y = y|A = a, Z = z)$ and $P(Z = z)$ are consistent across the real and the counterfactual world, and thus can be estimated using observational data from the original causal graph. This formula suffers from curse of dimension at Z but serves well as a pedagogical example. Some other popular causal techniques include *matching*, *covariate adjustment*, and *inverse probability (IP) weighting* (aka *propensity score*). More rigorous assumptions for a feasible causal estimation, such as positivity, can be found in [Miguel A. Hernán, 2018].

2.1.1 Individual treatment effect

Individual treatment effect (ITE) is gaining more and more popularity especially in patient-centric healthcare, the most fine grained one of which is defined as follows with a binary treatment.

$$ITE^{(1)} = E[Y^{A=1}|Z = z] - E[Y^{A=0}|Z = z] \tag{2.3}$$

where Z also includes patient-specific confounders. However, for a particular instance, only one of two terms in Equation 2.3 is available, thus ITE is often cast as a missing problem. The estimation of ITE is evident under the *potential outcome framework* [Miguel A. Hernán, 2018].

$$E[Y^a|Z = z] = E[Y^a|A, Z = z] = E[Y|A = a, Z = z] \tag{2.4}$$

holds when $Y^a \perp\!\!\!\perp A|Z$ holds, which means if Z is conditioned, the value of Y_a is not affected by value of A. This may first seem contradictory, but can be understood easier in a perspective of *structural equation modeling*. Suppose

in the underlying generative process, the equation for the outcome is $Y = A+Z+\epsilon$, where ϵ is caused by unobserved factors that are uncorrelated with A. Then under the intervention of $A = a$, the equation becomes $Y = a+Z+\epsilon$. At this stage, Y can still be affected by A in an observational study, since A and Z are correlated However, once Z is conditioned, $E[Y = a + z + \epsilon|Z = z] = E[Y = a + z + \epsilon|Z = z, A]$.

Hence the estimation of ITE turns out to be easier than that of ATE, as it only requires a regression surface of $E[Y|A = a, Z = z]$ but not other components. Nevertheless, as we will see later, this may not be true in some cases where the estimation of ITE is impossible while that of ATE works smoothly.

2.1.2 Time-varying treatments

New issues arise after time-varying treatments are involved. Previously, we only discuss causality that ignores time variation, assuming ATE remains unchanged regardless of when an outcome is measured after treatment. In this subsection, we extend our discussion to a more complicated scenario, where a patient receives a sequence of treatments and reports a final outcome at the end.

Many aforementioned concepts need generalizing to adapt to this new scenario. The definition of ATE for time-varying treatments can be extended to the difference between two treatment strategies, for example, between "always treat" and "never treat" strategies. The number of possible paired strategies increases exponentially with the number of treatments.

Accordingly, there needs a generalization of NUC for time-varying treatments, a concept of *sequential conditional exchangeability* [Miguel A. Hernán, 2018], where no direct unmeasured factor is taken into account to assign treatments, which is formally defined as follows.

$$Y^{\bar{a}_k} \perp\!\!\!\perp A_k|\bar{A}_{k-1}, \bar{Z}_k, \text{ for all strategies k=0,1...K} \tag{2.5}$$

However, as we will see later, even with NUC satisfied, traditional causal techniques still give a biased estimation of ATE when comparing treatment strategies.

Figure 2.1d demonstrates the epitome of cases for time-varying treatments, where only two treatments A_1, A_2 are taken into account. The problem with traditional causal techniques is that, when conditioned on a collider Z_1, A_1 and U are correlated. A backdoor path is then opened, $A_1 \rightarrow Z_1 \leftarrow U \rightarrow Y$. This always happens when there exist arrows from A_1 to Z_1 and from Z_1 to A_2, therefore named *treatment-confounder feedback* [Miguel A. Hernán,

2018]. Fortunately, this issue can be tackled by generalized causal techniques, including *g-formula* and *g-estimation*.

We illuminate the idea behind g-methods by a proof of *g-formula*. As before, a modified causal graph is first obtained and shown in Figure 2.1e, upon which the potential outcome can be derived.

$$P[Y^{a_1,a_2} = y] = P(Y = y|A_1 = a_1, A_2 = a_2) \tag{2.6}$$

$$= \sum_{u,z_1} \frac{P(Y = y, U = u, Z_1 = z_1, A_1 = a_1, A_2 = a_2)}{P(A_1 = a_1, A_2 = a_2)} \tag{2.7}$$

$$= \sum_{u,z_1} P(Y = y, U = u, A_1 = a_1, Z_1 = z_1, A_2 = a_2) \tag{2.8}$$

$$= \sum_{u,z_1} P(U = u)P(Z_1 = z_1|U = u, A_1 = a_1) \\ P(Y = y|U = u, A_1 = a_1, Z_1 = z_1, A_2 = a_2) \tag{2.9}$$

which Equation 2.8 and 2.9 hold since $P(A_1 = a_1) = P(A_1 = a_1|U = u) = 1$ and $P(A_2 = a_2) = P(A_2 = a_2|U = u, A_1 = a_1, Z_1 = z_1) = 1$ in the counterfactual world. Now coming back to the observational world, terms involving U is incalculable since U is unobserved, but U can be marginalized out under the new NUC assumption.

$$P[Y^{a_1,a_2} = y]$$

$$= \sum_{u,z_1} P(u)P(z_1|u, a_1)P(y|u, a_1, z_1, a_2) \tag{2.10}$$

$$= \sum_{u,z_1} P(u)P(z_1|u, a_1)P(y|u, a_1, z_1, a_2)\frac{P(a_1|u)P(a_2|u, a_1, z_1)}{P(a_1)P(a_2|a_1, z_1)} \tag{2.11}$$

$$= \sum_{u,z_1} \frac{P(u, a_1, z_1, a_2)}{P(a_1)P(a_2|a_1, z_1)}P(y|u, a_1, z_1, a_2) \tag{2.12}$$

$$= \sum_{u,z_1} \frac{P(a_1)P(z_1|a_1)P(a_2|a_1, z_1)P(u|a_1, z_1, a_2)}{P(a_1)P(a_2|a_1, z_1)}P(y|u, a_1, z_1, a_2) \tag{2.13}$$

$$= \sum_{u,z_1} P(z_1|a_1)P(u|a_1, z_1, a_2)P(y|u, a_1, z_1, a_2) \tag{2.14}$$

$$= \sum_{z_1} P(z_1|a_1)P(y|a_1, z_1, a_2) \tag{2.15}$$

where NUC applies in Equation 2.11 so that $P(a_1|u) = P(a_1)$ and $P(a_2|u, a_1, z_1) = P(a_2|a_1, z_1)$. An intuitive way to understand is that g-formula replaces $P(z_1)$ in traditional methods with $P(z_1|a_1)$, which equivalently replaces Z_1 with $Z_1^{a_1}$ in the causal graph.

ATE turns out to be identifiable with time-varying treatments, whereas $ITE^{(1)}$ does not have fortune to fall into the same category. $ITE^{(1)}$ of a strategy of treatments is, to the best of our knowledge, impossible without a stronger assumption, which can be seen by examining $Y^{a_1,a_2} \not\perp\!\!\!\perp A_1, A_2|Z_1$. However, if we enlarge the granularity of ITE, by splitting covariates into $\bar{Z}$ and patient-specific time-invariant factor L, as shown in Figure 2.1f, the following estimation is feasible.

$$ITE^{(2)} = E[Y^{\bar{a}_k}|L] - E[Y^{\bar{a}'_k}|L] \tag{2.16}$$

since L can be considered to exist before all treatments and be fixed, thus not causing any treatment-confounder feedback.

In this thesis we turn our attention to ITE of the most recent treatment, given a history of previous treatments. That is to say, instead of $E[Y^{\bar{a}_k}] - E[Y^{\bar{a}'_k}]$, we study

$$ITE^{(3)} = E[Y^{a_k}|\bar{A}_{k-1}, \bar{Z}_k] - E[Y^{a'_k}|\bar{A}_{k-1}, \bar{Z}_k] \tag{2.17}$$

$$= E[Y|\bar{a}_{k-1}, a_k, \bar{z}_k] - E[Y|\bar{a}_{k-1}, a'_k, \bar{z}_k] \tag{2.18}$$

which obviously holds in light of the new NUC assumption in Equation 2.5. To explain the discrepancy among different ITEs, the one we adopt gives a difference in treatment effects between two current latest treatments a_k and a'_k among patients who took $\bar{a}_{k-1}$ before. One interpretation for $ITE^{(3)}$ is that the effect of the most recent treatment a^k, also known as the short-term effect of $\bar{a}_k$ [Miguel A. Hernán, 2018]. It is also worth noting that when estimating $E[Y^{\bar{a}_k}]$, the model implicitly assumes *new-user designs* where patients had not used treatment in the past.

We will focus on the most recent treatment in following subsections, since previous treatments now serve a same role as past covariates $\bar{Z}_k$, and do not pose any new conceptual difficulty.

2.1.3 Time-varying outcomes

A treatment usually introduces a long duration of effect, which can not be fully described by a time-fixed value. For example, Zeevi et al. [2015] describes as a scalar area a time-varying glycemic response caused by a food intake, losing details in progression of response. Therefore there comes the need to analyze time-varying outcomes.

Looking at Figure 2.1g which exemplifies the case of time-varying outcomes within three time steps, estimating $E[Y^a_t]$ is very straightforward. But it may appear that when estimating Y^a_{t+1} at the next time step, conditioning on Y_t opens a backdoor path $A \to Y_t \leftarrow U_t \to U_{t+1} \to Y_{t+1}$. However,

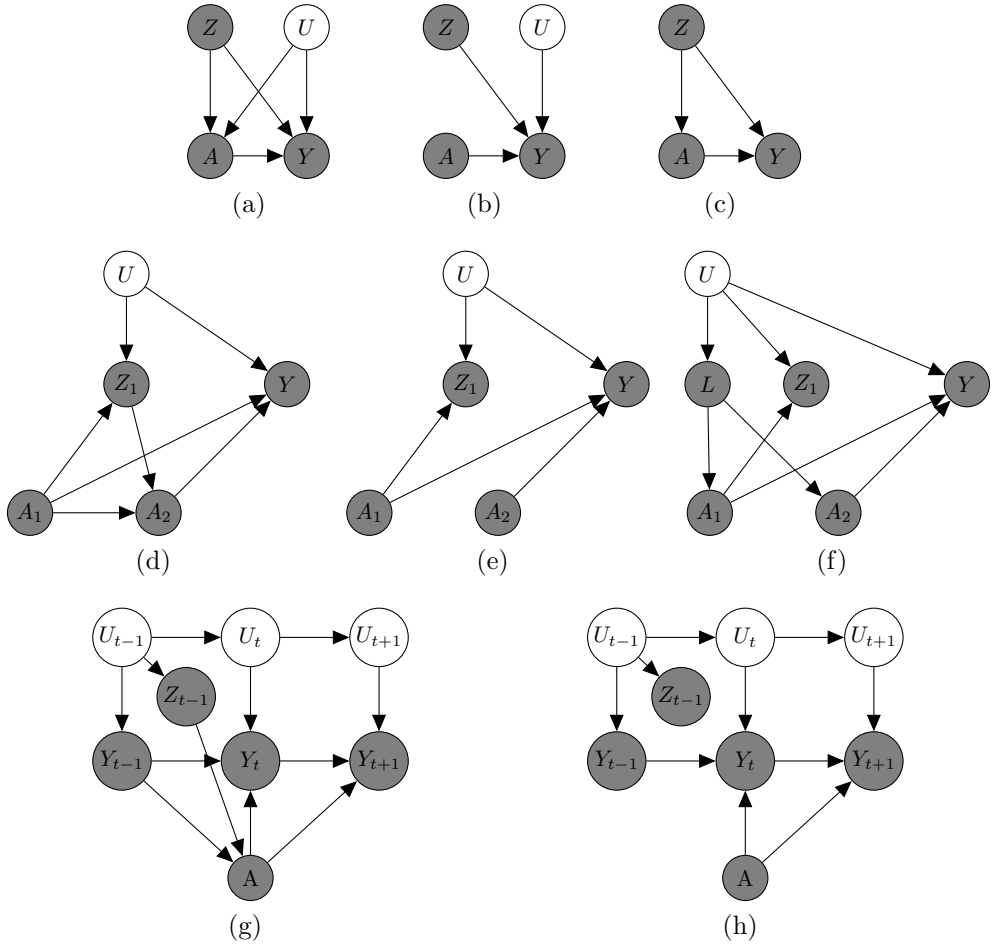


Figure 2.1: Causal graphs for different settings

this is not true, since Y_t is actually an unknown future variable, and what is conditioned for Y_{t+1} is Y_t^a, for which $Y_t^a \perp\!\!\!\perp A | Y_{t=1}$ holds.

$$E[Y_t^a | Y_{t-1}, Z_{t-1}] = E[Y_t | A = a, Y_{t-1}, Z_{t-1}] \tag{2.19}$$

$$E[Y_{t+1}^a | Y_t^a, Y_{t-1}, Z_{t-1}] = E[Y_t | A = a, Y_t^a, Y_{t-1}, Z_{t-1}] \tag{2.20}$$

We ignore Z_t, Z_{t+1} because they are future variables which by no means will affect the current treatment A. Generalizing above formulas to an infinite time series together with past treatments, we have

$$E[\bar{Y}_{\geq t}^{\bar{a}_{<t}, a_t} | \bar{a}_{<t}, \bar{Y}_{<t}, \bar{Z}_{<t}] = E[\bar{Y}_{\geq t} | \bar{a}_{\leq t}, \bar{Y}_{<t}, \bar{Z}_{<t}] \tag{2.21}$$

2.1.4 Causal applications

Dahabreh et al. [2012] conducts a systematic comparison between the results of observational study and RCTs on therapeutic interventions for acute coronary syndromes, and finds a high consistency between them. However, Gordon et al. [2018] reports an inconsistency in the context of advertising campaigns. Thus, results from observational studies should be applied with caution

Compared to the single outcome scenario, related work on time-varying outcomes is much fewer. Brand and Xie [2007] discusses cases of time-varying treatments and time-varying outcomes, however, with a special dichotomous irreversible treatment, e.g., disability, which avoids treatment-confounder feedback and leads to a simpler solution. Schulam and Saria [2017] proposes a stronger continuous NUC assumption to ensure the unbiased estimation of $ITE^{(1)}$ over a treatment strategy. Soleimani et al. [2017] captures the dynamics in treatment-response trajectories by convolving a dose function with a parametric impulse response function. Lim [2018] leverages the memory mechanism in RNN to learn generalized propensity scores and a second sequence-to-sequence network to make multiple-step prediction of $ITE^{(2)}$.

Apart from what have been discussed, another group of causal methods that capture an additive unobserved effect caused by time-invariant confounders is *fixed-effects models* (FE) [Wooldridge, 2010], which has been widely employed in fields of social science. Typical study objects of FE are panel or cohort data.

2.2 Errors-in-variables

Massive data for model training are usually collected without a careful inspection or with an inevitable instrumental error, thus being contaminated to some extent. While most regression methods permit an unbiased homoscedastic error in the dependent variable, errors in independent variables are typically omitted. However, the former only induces additional variability in estimation in either linear or nonlinear models [Carroll et al., 2006, Chapter 15] whereas the latter is more destructive than practitioners think– leading to a biased estimation, which can not be remedied with even infinite samples. Models that take into account measurement errors in independent variables are called *errors-in-variables* models.

We will first introduce common formulations and consequences of mismeasurement error, and then proceed to various corrective solutions, mostly in linear models. Some general advice regarding important decisions in EIV

modeling is given below, and readers may want to come back to this after reading this section. Other than linear regression, EIV modeling almost always requires auxiliary information or data in order to correct bias in estimation. Plug-in correction is suitable for problems that have an analytical solution. To absorb information from instrumental variables or repeated measurements is not straightforward, especially in nonlinear models. When no additional data is available, Bayesian EIV approach is by far the most powerful and flexible one in coping with non-identifiability, by applying additional information as distributional assumptions.

In this discussion, some important topics are not covered, such as bounds of coefficients (aka sensitivity analysis) and EIV on panel data. An extensive treatment to EIV can be found in Carroll et al. [2006] Schennach [2012] Gustafson [2004] Chen et al. [2007].

2.2.1 Consequences and formulations of mismeasurement

An important category of mismeasurement is *non-differential error*, which claims that measurement error is independent of the dependent variable. In other words, mismeasurement does not contribute any new information to the regression. Error-prone X is then termed a *surrogate* or *proxy* for its true variable X^*. Only non-differential error will be consider within the scope of this thesis, because when the error is not non-differential, i.e., differential, it becomes more difficult or impossible to eliminate the bias.

2.2.1.1 Classical additive error

. The most important type of mismeasurement is *classical measurement error*, where measurement error is independent of its latent variable. In this thesis, when mentioning classical error, we refer to the most commonly seen *classical additive error*, which takes a formulation as follows.

$$X = X^* + \Delta X \tag{2.22}$$

where ΔX is an independent unbiased additive error such that $\Delta X \perp X^*$, $E[\Delta X] = 0$ and $Var(\Delta X) = \sigma_{\Delta X}$. Classical error induces a biased and inconsistent estimator even in the simplest case of simple linear regression.

$$Y = \beta_0 + \beta X^* + \epsilon \tag{2.23}$$

where we assume $E[\epsilon|X] = E[\epsilon] = 0$ and $Var(\epsilon) = \sigma_\epsilon$, i.e., ϵ is an independent random noise. Because of measurement error, X is observed instead of X^*.

$$Y = \beta_0 + \beta X + \eta \tag{2.24}$$

where $\eta = \epsilon - \beta\Delta X$, which is correlated to the actual regressor X. The according estimation becomes,

$$\begin{aligned} Cov(Y, X) &= E[(Y - \bar{Y})(X - \bar{X})] && (2.25)\\ &= E[(\beta X + \epsilon - \beta\Delta X - \beta\bar{X})(X - \bar{X})] && (2.26)\\ &= E[(\beta X - \beta\Delta X - \beta\bar{X})(X - \bar{X})] && (2.27)\\ &= \beta E[(X - \bar{X})(X - \bar{X}) - \Delta X(X - \bar{X})] && (2.28)\\ &= \beta E[(X - \bar{X})^2 - \Delta X^2] && (2.29)\\ &= \beta(Cov(X, X) + \sigma_{\Delta X}) && (2.30) \end{aligned}$$

Therefore, a naive estimator $\hat{\beta}_{\Delta X}$ is biased.

$$\hat{\beta} = Cov(Y, X^*)/Cov(X^*, X^*) \mapsto \beta \quad (2.31)$$

$$\hat{\beta}_{\Delta X} = Cov(Y, X)/Cov(X, X) \not\mapsto \beta \quad (2.32)$$

Notice that $Cov(Y, X) = Cov(Y, X^*)$ and $Cov(X, X) = Cov(X^*, X^*) + \sigma_{\Delta X}$. Thus,

$$\hat{\beta}_{\Delta X} = \frac{Cov(Y, X^*)}{Cov(X^*, X^*) + \sigma_{\Delta X}} = \frac{Cov(X^*, X^*) + \sigma_{\Delta X}}{Cov(X^*, X^*)}\hat{\beta} \quad (2.33)$$

$$\hat{\beta} = \frac{Cov(X^*, X^*)}{Cov(X^*, X^*) + \sigma_{\Delta X}}\hat{\beta}_{\Delta X} = \Gamma\hat{\beta}_{\Delta X} \quad (2.34)$$

where Γ is called a *reliability ratio*, This kind of bias is called *attenuation*, indicating the magnitude of coefficients shrink towards zero. As we can see, a larger error (i.e., a larger $\sigma_{\Delta X}$) results in a smaller reliability ratio.

Another similar error, *mean-independence error* relaxes the requirement of independence between X^* and ΔX, only requiring $E[\Delta X|X^*] = 0$, which allows a heteroscedastic ΔX.

2.2.1.2 Berkson additive error

. While classical error causes detrimental effects on simple linear regression, another similar error, *Berkson error*, does not. Berkson error takes a following form, with places of X^* and X switched.

$$X^* = X + \Delta X \quad (2.35)$$

Typically, if X^* can be remeasured, then the error is classical instead of Berkson. The key property in Berkson error is that observed error-prone X is independent of ΔX. Therefore fitting a linear regression of Y directly on X gives an unbiased coefficient estimation. However, error always has negative influence. X^* with Berkson error has a larger variance, which reduces statistical power, e.g., in hypothesis testing [Carroll et al., 2006, Chapter 1].

2.2.1.3 Classical multiplicative error

Another common type of error is *multiplicative error*.

$$X = \gamma X^* \tag{2.36}$$

A natural impulse is to apply logarithm transformation so that it turns into classical additive error. However, this does not work since substituting $\log X$ for X leads to another distinct coefficient $\beta_{\log}$, instead of β.

$$Y = \beta_{0,\log} + \beta_{\log} \log X^* + \epsilon \tag{2.37}$$

Hwang [1986] proposes a consistent estimator for multiplicative cases.

Nevertheless, if $\beta_{\log}$ is of main interest, then employing techniques for classical additive error is plausible. One noteworthy point is the distribution of γ, which is typically assumed to be LogNormal distribution.

2.2.2 Plug-in correction

After deriving the reliability ratio Γ in simple linear regression, it is straightforward to correct the bias by multiplying $\hat{\beta}_{\Delta X}$ with Γ. This may seem impractical at first sight, but the knowledge about $\sigma_{\Delta X}$ is often available, for example, when the mismeasurement is caused by a specific machine whose error variance is known, similar studies provide external validation data where X and X^* are both known, or there are partial replicates of X. Once $\sigma_{\Delta X}$ is known, $Cov(X^*, X^*) = Cov(X, X) - \sigma_{\Delta X}$. Then Γ can be calculated.

For other models, as long as an analytical estimation can be derived, then all unknown components can be sought or estimated from external sources to serve as plug-in components to restore the true coefficient.

Some common sources of additional information that can help correct the bias caused by EIV are listed below.

1. $\sigma_{\Delta X}$;
2. Distribution of ΔX, aka the error model;
3. Distribution of X^*, aka exposure model;
4. External validity data;
5. Partial internal validity data;
6. IV (Subsection 2.2.3);
7. Partial replicates.

2.2.3 Instrumental variables

Instrumental Variables (IV) is considered the second most important technique to Least Squares in econometrics. IV provides a solution to regression where independent variables are correlated with the residual error term of the dependent variable, for which traditional least squares estimator produces a biased result. This technique can be applied for many issues, including EIV and causality.

IV also refers to variables V that satisfy following requirements.

$$V \not\perp X^*,\ V \perp Y|X^*,\ V \perp \Delta X \tag{2.38}$$

To see why V can help with error-prone X, let us enunciate by an example where $X^* = \alpha V + W$ where W is the component in X^* that is not related to V.

$$\begin{aligned} Cov(X,V) &= Cov(\alpha V + W + \Delta X, V) && (2.39)\\ &= Cov(\alpha V + W, V) && (2.40)\\ &= Cov(X^*, V) && (2.41)\\ Cov(Y,V) &= Cov(\beta(\alpha V + W) + \epsilon, V) && (2.42)\\ &= Cov(\beta(\alpha V + W), V) && (2.43)\\ &= \beta Cov(\alpha V + W, V) && (2.44)\\ &= \beta Cov(X^*, V) && (2.45) \end{aligned}$$

Thus we are able to obtain $\hat{\beta} = Cov(Y,V)/Cov(X,V)$. This estimation holds given an arbitrary relation between X^* and V, not limited to a linear one.

However, choosing a valid V remains a matter of art in practice. If any of requirements over V is violated, this estimator will be biased. One possible choice for V is an independent replicate measurement of X^*, which shows an important connection that IV is a relaxed requirement of an independent replicate measurement.

2.2.4 EIV without side information

As we see in previous subsections, a naive estimator presents a biased result under classical error, and a modified estimator typically requires auxiliary data or information. A natural question to ask is whether it is possible to correct the bias without any side information. This question is related to a notion of *identifiability*, the lack of which indicates an essential component is missing in the modeling. This subsection investigates identifiability of EIV modeling and *methods of moments* without side information.

2.2.4.1 Identifiability

In statistics, *identifiability* requires an one-to-one mapping between the statistical model $P(Y|\Theta)$ and its parameter Θ. That is to say, no two distinct parameters lead to the same probability distribution for observed data Y, in which Y here includes both the outcome and regressors. The opposite direction also holds–with an infinite amount of observed data Y, we can reconstruct the true distribution by the strong law of large numbers, which corresponds to a unique parameter. This provides a theoretical foundation for identification of the true model parameter from observed data.

A simple example for non-identifiability is linear regression with a singular design matrix, i.e., its underlying data-generating process produces linear dependent regressors. Assume two different parameters Θ_1, Θ_2 share one density function $P(Y)$, in a sense of non-zero measure. Then their likelihood are equal.

$$\mathcal{L}(\Theta_1) = \sum_i^N P(y_i|\Theta_1) = \sum_i^N P(y_i|\Theta_2) = \mathcal{L}(\Theta_2) \tag{2.46}$$

This is a special case of *multimodal* likelihood functions, in which MLE has more than one solution.

One needs to stay cautious even when the model is identifiable, because the identification in practice may not always be on a par with its theoretical counterpart. Given an identifiable model, it can still be close to unidentifiability, especially in nonlinear models, for example, when $\mathcal{L}(\Theta_1)$ is merely slightly smaller than $\mathcal{L}(\Theta_2)$. Performance of current optimization methods heavily rely on a good initial point, which means it is unstable and likely to converge to a nearby local maximum. Carroll et al. [2006, Chapter 8] also warns practitioners not to be optimistic about technical identifiability of models.

There are several ways to cope with nonidentifiability. The first one is to impose constraints on model structure. Adams et al. [2019] ensures identifiability by assuming strict underreporting, i.e. $P(X = 1|X^* = 0) = 0$, and choosing a special family of restricted models in a problem of exposure misclassification. The Bayesian EIV approach covered in Subsection 2.2.5 is another general cure for nonidentifiable models, by applying informative priors on nonidentified nuisance parameters.

2.2.4.2 Methods of moments

In the seminal work [Geary, 1941], a consistent EIV estimation can be obtained in linear models using moments (or cumulants) of third or higher

order, under mild assumptions other than knowing additional error distribution or having auxiliary data or IV. Pal [1980] generalizes this idea and establishes more possible moment-based consistent estimators.

The main idea of methods of moments is that, estimation problem is formulated as a system of equations, where each equation is comprised of moments and parameters. Once the number of equations is equal or larger than the number of parameters, then the whole system can be uniquely solved. Take simple regression in Equation 2.23 as an example. If only first and second moments are allowed, then the number of equations is not enough to lead to a unique solution, i.e. 5 equations and 6 parameters, $E[X^*], E[X^{*2}], \beta, \beta_0, \sigma_\epsilon$ and $\sigma_{\Delta X}$.

$$E[X] = E[X^*] \tag{2.47}$$
$$E[Y] = \beta_0 + \beta E[X^*] \tag{2.48}$$
$$E[X^2] = E[X^{*2}] + \sigma_{\Delta X} \tag{2.49}$$
$$E[Y^2] = \beta_0^2 + \beta^2 E[X^{*2}] + 2\beta_0\beta E[X^*] + \sigma_\epsilon \tag{2.50}$$
$$E[XY] = \beta_0 E[X^*] + \beta E[X^{*2}] \tag{2.51}$$

If higher order moments are considered, then the system is solvable with 7 equations and 7 parameters.

$$E[(X - \bar{X})^3] = E[(X^* - \bar{X^*})^3] \tag{2.52}$$
$$E[(Y - \bar{Y})^3] = \beta^3 E[(X^* - \bar{X^*})^3] \tag{2.53}$$

The coefficient of our interest can be rewritten as a formula of moments. One possible one is shown below for simple regression.

$$\hat{\beta}_M = \sqrt[3]{\frac{E[(Y - \bar{Y})^3]}{E[(X - \bar{X})^3]}} \tag{2.54}$$

This estimator is consistent as long as $E[(X^* - \bar{X^*})^3] \neq 0$, which means the distribution of X^* is not symmetric. In cases where third order moment of X^* is zero, other estimators using fourth order moments are possible, as long as $\kappa_4(X^*) \neq 0$.

This may already remind some readers of a well-known and unique property of Normal distribution–third and higher order cumulants are zeros. Reiersøl [1950] formally proves that EIV is not identifiable in simple regression if all variables, including independent and dependent variables and their error terms, are normally distributed. Schennach et al. [2007] makes a further step to conclude

$$g(X^*) = a + b\ln(e^{cX^*} + d) \tag{2.55}$$

is the only form in $Y = g(X^*) + \Delta Y$ that can not be identified under conditions of monotonous and smooth $g(\cdot)$ and non-vanishing characteristic functions of error terms, which includes Reiersøl [1950] as a special case, i.e. when $d = 0$.

2.2.5 Bayesian methods

In cases where neither an analytical solution nor auxiliary data are available, which is common in practice, additional assumptions are required in order to complete EIV modeling. One important approach that assumes a distributional assumption is to delineate a joint distribution for observed variables as if unknown variables such as X^* were available. In previous subsections, we assume X^* is fixed, which is called *functional* modeling. For now, we devote the discussion to a *structural* manner where X^* is assumed to be a random variable with a density function $P(X^*)$.

$$P(Y, X) = \int P(Y, X, X^*)\,\mathrm{d}x^* \tag{2.56}$$

$$= \int P(Y|X, X^*)P(X|X^*)P(X^*)\,\mathrm{d}x^* \tag{2.57}$$

$$= \int P(Y|X^*)P(X|X^*)P(X^*)\,\mathrm{d}x^* \tag{2.58}$$

where $P(Y|X^*)$ is named *outcome model*, $P(X|X^*)$ is *error model*, and $P(X^*)$ is *exposure model* which is a term borrowed from epidemiology. Different structure of an outcome model can be freely adopted depending on the target application. An error model often chooses Normal distribution. The most problematic part is the decision for exposure model, which usually depends on application context.

This approach also provides another perspective to understand the discrepancy between classical error and Berkson error.

$$P(Y, X) = \int P(Y, X, X^*)\,\mathrm{d}x^* \tag{2.59}$$

$$= \int P(Y|X, X^*)P(X^*|X)P(X)\,\mathrm{d}x^* \tag{2.60}$$

$$= \int P(Y|X^*)P(X^*|X)P(X)\,\mathrm{d}x^* \tag{2.61}$$

$$P(Y|X) = \int P(Y|X^*)P(X^*|X)\,\mathrm{d}x^* \tag{2.62}$$

We obviate the exposure model by modeling a conditional distribution instead of a joint one.

There are two major approaches of inference.

$$\mathcal{L}(\Theta) = \sum_{i}^{N} P(y_i, x_i) \tag{2.63}$$

where Θ is the set of parameters for the model. After specifying a likelihood function $\mathcal{L}(\Theta)$ of parameters including unknown X^* using the joint distribution and observed data, perform *Maximum Likelihood Estimation* (MLE), which is a point estimation. Typical optimization techniques include gradient-based methods and Newton methods. By further instilling human knowledge into parameters via prior distributions, this method develops into a Bayesian method.

$$P(\Theta|D) = \frac{P(\Theta)\mathcal{L}(\Theta)}{P(D)} \tag{2.64}$$

$$\propto P(\Theta)\mathcal{L}(\Theta) \tag{2.65}$$

where D stands for the set of observed data, and $P(\Theta)$ a prior distribution of parameters. Then the optimization of posterior distribution $P(\Theta|D)$ are usually conducted by *Markov Chain Monte Carlo* (MCMC).

Although Bayesian EIV approach seemingly works well, same as other structural models, it suffers from *non-identifiability* when key information is missing, which is a notion discussed in Subsection 2.2.4.1. Nevertheless, Bayesian priors make flexible the assimilation of additional information. Compared to plug-in correction, where an exact point estimation of unknown corrective components is required, Bayesian EIV allows only a rough distribution of unknown parameters, which is more plausible in practice. Gustafson [2004] points out, with proper priors, Bayesian EIV works reasonably well even in non-identifiable models.

Despite being able to tackle a wider range of problems, Bayesian EIV methodology has its own disadvantages. First of all, it is sensitive to misspecification. Determining a proper distribution structure for $P(X^*)$ is an art, and any misspecification is likely to lead to untrustworthy result. Secondly, simulation-based methods are computationally heavy, especially for high-dimensional parameters.

2.3 Gaussian Processes

GPs [Rasmussen, 2004] are non-parametric regression models with closed-form posterior estimation, which they inherit from the Normal distribution

by assuming that all training and test data follow a joint multivariate Normal distribution. GPs can also be extended for classification Rasmussen [2004]. In the context of our thesis, GPs are used to construct a time series, i.e., to regress outcomes Y on time T.

$$Y(t) \sim \mathcal{GP}(0, k(t, t'|\theta)),$$

where θ are parameters associated with the *kernel* function $k(x, x'|\theta)$, which produces valid covariance matrix with desired properties, e.g., smoothness. Kernels will be introduced in more details later.

A chosen kernel puts a constraint or prior on the function space, and conditioning on training data selects most likely functions as its posterior. Leveraging the good properties of Normal distribution, posterior distribution of GPs on testing data can be derived as

$$\begin{aligned} Y(t)|\mathbf{Y}_n &\sim N(\mu_*, \Sigma_*), \quad \text{where} \\ \mu_* &= k(\mathbf{t}_n, t)^T K(\mathbf{t}_n, \mathbf{t}_n)^{-1} \mathcal{S}_n, \quad \text{and} \\ \Sigma_* &= k(t, t) - k(\mathbf{t}_n, t)^T K(\mathbf{t}_n, \mathbf{t}_n)^{-1} k(\mathbf{t}_n, t). \end{aligned}$$

where $\mathbf{t}_n$ and $\mathbf{Y}_n$ are training data. We refer the reader to Rasmussen [2004] for more details about GPs.

As the kernel, the sum of the standard Squared Exponential (SE) and constant kernels is used in this thesis. A constant kernel assumes every pair of input share a constant covariance, which is useful for extrapolation so that outcomes with an input outside the training input region take an average value instead of dropping to zero immediately.

$$k(x, x') = C \tag{2.66}$$

GPs with a SE kernel are blessed with desired smoothness, as two close inputs show a strong correlation, while distant ones are uncorrelated.

$$k(x, x') = \sigma^2 exp(-\frac{(x - x')^2}{2l^2}) \tag{2.67}$$

A sum of two kernels presents an OR operation.

To speed up computation where a normal GP has a $\mathcal{O}(n^3)$ cost, we use a *sparse* GP [Rasmussen, 2004] instead of a full GP, which samples a small set of inducing points uniformly from $\mathbf{t_n}$ to achieve a low-rank approximation of $K(\mathbf{t_n}, \mathbf{t_n})$ and its inverse.

Chapter 3

Methods

In this section, we describe three major components of our model for personalized treatment-response trajectories: hierarchical parametric treatment responses, a counterfactual trend modeled by a Gaussian Process, and measurement error models. Throughout the section, we present the model in generic terms, but also outline the specific model that we use in Section 4 to estimate the impact of diet, recorded as nutrient contents of different meals, on continuous blood glucose measurements. Besides, we also derive closed form marginal increment of treatment response area for interpretability of our model.

Our model is fully Bayesian, yielding uncertainty estimates for all parameters, which is essential in scientific applications. Inference is done using Markov chain Monte Carlo (MCMC) with the state-of-the-art No-U-Turn (NUTS) sampler [Hoffman and Gelman, 2014] implemented in software PyMC3 [Salvatier et al., 2016].

3.1 An overview

A graph of our model for treatment-response trajectories is presented in Figure 3.1. We assume there are N patients, and a trajectory consisting of a time series of length G_n of the outcome (e.g. blood glucose) is observed for each individual:

$$\mathbf{y_n} = (y_{n1}, \ldots, y_{nG_n})^T, n = 1, \ldots, N.$$

These measurements have been taken at times

$$\tau_{\mathbf{n}} = (\tau_{n1}, \ldots, \tau_{nG_n})^T, n = 1, \ldots, N.$$

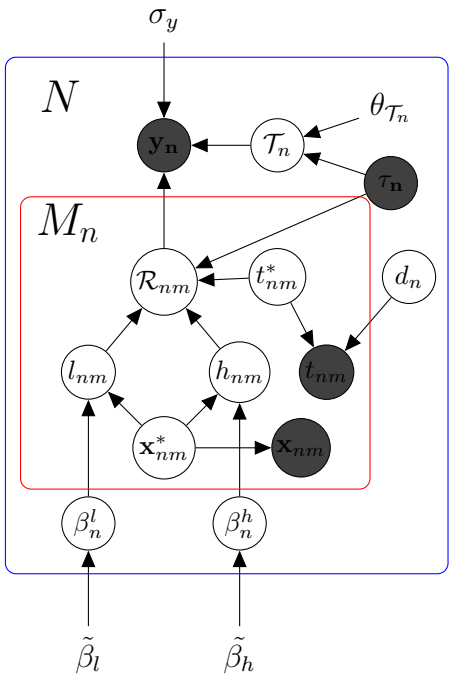


Figure 3.1: Model specification

Furthermore, each patient has M_n observed treatments (e.g. meals eaten), indexed by $m \in 1, \ldots, M_n$, where each treatment is characterized by P covariates:

$$\mathbf{x}_{nm} = (x_{nm1}, \ldots, x_{nmP})^T, \text{for all } m, n,$$

and the corresponding recorded treatment times are

$$\mathbf{t}_n = (t_{n1}, \ldots, t_{nM_n})^T, \text{for all } n.$$

Here, $\mathbf{x}_{nm}$ and t_{nm} are assumed to be noisy observations of the treatment covariates and timings, and their true unobserved values are denoted by $\mathbf{x}^*_{nm}$ and t^*_{nm}, respectively.

Outcome model: We model the observed outcome trajectory of individual n, $\mathbf{y}_n$, as

$$\mathbf{y}_n = \mathcal{T}_n + \sum_m \mathcal{R}_{nm} + \mathbf{e},$$

where $\mathcal{T}_n \in \mathbb{R}^{G_n}$ is a counterfactual trend (i.e. it describes the evolution of the outcome had the treatment not been taken), $\mathcal{R}_{nm} \in \mathbb{R}^{G_n}$ is the response to the mth treatment, and $\mathbf{e} = (e_1, \ldots, e_{G_n})^T$ is the vector of errors with $e_i \sim N(0, \sigma_y^2)$. We note that the sum of the trend and the responses can be viewed as a trajectory for a 'clean' outcome (omitted from Figure 3.1), of

which a version corrupted by Gaussian noise is observed. Additive response functions can be seen as a continuous extension to scalar *average treatment effect* (ATE) which is the expected difference of outcomes before and after treatment.

3.2 Response function

Response functions specify how the treatment affects the outcome over time, and they should be specified to suit the application at hand, by balancing between flexibility, interpretability, etc. For example, if interpretability is not needed and the amount of data is large, then non-parametric functions that automatically learn the shape of the response are attractive. On the other hand, parametric functions are a viable option when data are scarce. Furthermore, they are often interpretable, which is both valuable in itself but also allows using prior knowledge to further improve accuracy.

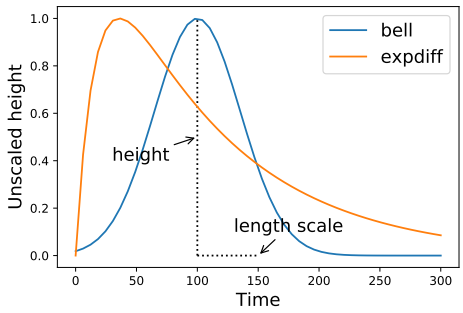


Figure 3.2: Two Response functions. The blue one is used in this paper, while the orange one is used in Schulam and Saria [2017].

For the application of modeling the impact of meals on blood glucose, considered in Section 4, we select the treatment response as a bell-shaped parametric function

$$\begin{aligned} \mathcal{R}_{nm} &:= f(\Delta_{nm}, h_{nm}, l_{nm}) \\ &:= h_{nm} \exp\left\{ \frac{-0.5(\Delta_{nm} - 3l_{nm})^2}{l_{nm}^2} \right\}, \end{aligned} \tag{3.1}$$

where a lag vector $\Delta_{nm} = \tau_n - t^*_{nm}$ is introduced to represent the time since a specific treatment. The shape of the response in Equation (3.1) is shown in Figure 3.2 and is determined by two parameters h_{nm} and l_{nm}, which have straightforward interpretations: h_{nm} is the height of the response, and l_{nm}

is the length-scale which is directly proportional to the 'width' or 'duration' of the response. The main challenge in this application is treatment data scarceness, with only about 10 meals on average for each patient. We made an attempt to use a more flexible three-parameter function in Schulam and Saria [2017] that allows skewed response progression which leads to an inferior result.

In applications it is often of interest to measure how the response depends on treatment covariates, and therefore we allow the parameters to depend on the covariates:

$$\begin{aligned} h_{nm} &= (\beta_n^h)^T \mathbf{x}_{nm}^*, \text{ and} \\ l_{nm} &= (\beta_n^l)^T \mathbf{x}_{nm}^*, \text{ for all } n, m. \end{aligned} \tag{3.2}$$

In Equation (3.2), the coefficient vectors $\beta_n^h, \beta_n^l \in \mathbb{R}^P$ represent the *personalized impact* of each of the P covariates on the height or width of the response for the nth individual.

To share information across individuals, we introduce a Bayesian hierarchical prior [Gelman et al., 2013]. Accordingly, we assume the personalized height and length-scale coefficients, β_n^h and β_n^l, come from common distributions:

$$\beta_n^h \sim N_P(\tilde{\beta}_h, \Sigma_h) \quad \text{and} \quad \beta_n^l \sim N_P(\tilde{\beta}_l, \Sigma_l).$$

A hyper prior is further placed on the mean parameters of these distributions:

$$\tilde{\beta}_h \sim N_P(\mathbf{0}, \tilde{\Sigma}_h) \quad \text{and} \quad \tilde{\beta}_l \sim N_P(\mathbf{0}, \tilde{\Sigma}_l)$$

The hierarchical prior introduces shrinkage and facilitates estimation of the personalized coefficients even with limited data.

3.3 Counterfactual trend

A counterfactual trend represents the outcome assuming no treatment has been taken. It has to be sufficiently flexible to handle any variation in the outcome that is not accounted for by the treatments. In this paper, we model the trend $\mathcal{T}_n(t)$ for individual n using a Gaussian Process (GP):

$$\mathcal{T}_n(t) \sim \mathcal{GP}(0, k(t, t'|\theta_{\mathcal{T}_n})),$$

where $\theta_{\mathcal{T}_n}$ are parameters associated with the kernel function $k(x, x'|\theta_{\mathcal{T}_n})$.

$$\begin{aligned}
\mathcal{S}_n &= \mathbf{y}_n - \sum_m \mathcal{R}_{nm}(\tau_n) \\
\mathcal{T}_n(t)|\mathcal{S}_n &\sim N(\mu_*, \Sigma_*), \quad \text{where} \\
\mu_* &= k(\tau_\mathbf{n}, t)^T K(\tau_\mathbf{n}, \tau_\mathbf{n})^{-1} \mathcal{S}_n, \quad \text{and} \\
\Sigma_* &= k(t,t) - k(\tau_\mathbf{n}, t)^T K(\tau_\mathbf{n}, \tau_\mathbf{n})^{-1} k(\tau_\mathbf{n}, t).
\end{aligned}$$

where $\mathcal{S}_n$ is the residual of the outcome after subtracting the impact of the treatment responses.

As the kernel, we use the sum of the standard Squared Exponential (SE) and constant kernels. To speed up computation, we use a sparse GP, which samples a small set (around 10%) of inducing points uniformly from $\tau_\mathbf{n}$. We refer the reader to Section 2.3 for more details about GPs.

3.4 Measurement models

Measurement models describe error in observations. With self-reported data both covariates and the timing of a treatment may be uncertain. To account for the uncertainty in treatment timing, we assume:

$$t_{nm} \sim N(t^*_{nm} + d_n, (\sigma^t_n)^2), \quad \text{for all } n, m.$$

In words, the observed time t_{nm} is obtained from the true time t^*_{nm} by shifting it with a bias term d_n, and adding Gaussian noise. The bias term represents the habits of different individuals in reporting treatments. For example, in the blood glucose application in Section 4, some individuals may systematically report their meal after eating, while others may do this before eating.

Different models are possible for the treatment covariates, depending on the assumptions and data available [Gustafson, 2004]. Here we assume a simple perturbation on the *amount* of treatment:

$$\begin{aligned}
x_{nm} &= x^*_{nm}\delta_{nm}, \quad \text{where} \\
\delta_{nm} &\sim \mathit{Lognormal}(0, \sigma_x^2), \quad \text{for all } n, m.
\end{aligned} \tag{3.3}$$

The coefficient δ_{nm} represents the error for the mth treatment of nth individual. Intuition in the blood glucose application is that users are able to report correctly what they have eaten, but not how much. While more complicated models can be justified, the model in Equation (3.3) has the benefit that we can train it with little data.

Estimating t^*_{nm} is straightforward as it only shifts the response, but does not change its shape. However, estimating x^*_{nm} is more complicated, and requires assuming that the counterfactual trend is sufficiently regularized. Otherwise the trend could easily compensate for the perturbation. We solve this by encouraging a large length-scale for the squared exponential kernel in the prior.

3.5 Marginal increment of treatment response area

For simplicity, we focus on a single individual and drop the unnecessary indexing in the notation. The area A is proportional to length-scale l and height h of the response. Hence

$$A = \lambda h l \tag{3.4}$$

for some constant λ (knowing the shape of the response, solving for λ analytically is straightforward). Denote the amount of one covariate, e.g. sugar, in the mth meal by x_{mi} where $i \in \{1, 2, ..., P\}$. Now the length-scale l depends on x through

$$l_m(x_{mi}) = g(y^l_m) = g(\beta^l_i x_{mi} + c^l_m) \tag{3.5}$$

where g is the *softplus* function and c^l_m comprises the other parts of the linear predictor that do not depend on the sugar x_{mi}. Similarly, the height h depends on x through

$$h_m(x_{mi}) = g(y^l_m) = g(\beta^h_i x_{mi} + c^h_m) \tag{3.6}$$

We want to know how area A_m changes if we change the amount of sugar x_{mi} by one unit.

$$\frac{dA_m}{dx_{mi}} = \lambda \frac{dl_m}{dx_{mi}} h_m + \lambda l_m \frac{dh_m}{dx_{mi}} \tag{3.7}$$

$$= \lambda \frac{dl_m}{dy^l_m} \frac{dy^l_m}{dx_{mi}} h_m + \lambda \frac{dh_m}{dy^h_m} \frac{dy^h_m}{dx_{mi}} l_m \tag{3.8}$$

$$= \lambda (1 + e^{-y^l_m})^{-1} \beta^l_i h_m + \lambda (1 + e^{-y^h_m})^{-1} \beta^h_i l_m \tag{3.9}$$

By replacing x_m with an average meal, we have

$$\frac{dA}{dx_i} = \lambda (1 + e^{-\bar{y^l}})^{-1} \beta^l_i \bar{h} + \lambda (1 + e^{-\bar{y^h}})^{-1} \beta^h_i \bar{l} \tag{3.10}$$

$$= \lambda (1 + e^{-(\beta^l)^T \bar{\mathbf{x}}})^{-1} \beta^l_i \bar{h} + \lambda (1 + e^{-(\beta^h)^T \bar{\mathbf{x}}})^{-1} \beta^h_i \bar{l} \tag{3.11}$$

Chapter 4

Experiments

In this chapter, we use our model to analyze a real-world dataset, provided by obesity research unit at Obesity Research Unit at the University of Helsinki, Finland. Throughout, we compare four models, in an increasing order of complexity (later models include the previous as special cases):

- $\mathcal{M}_{ind}$: Separate models for individuals, no EIV.
- $\mathcal{M}_{hier}$: Model with the hierarchical prior for the responses to share information across individuals.
- $\mathcal{M}_{hier+time}$: Time uncertainty included.
- $\mathcal{M}_{hier+time+cov}$: Uncertainty in covariates included.

4.1 Dataset

The data contains blood glucose values collected by a portable continuous glucose monitoring system every few minutes and user-reported daily diet records for 13 non-diabetic individuals across three days. The visualization of the data (and results) for one individual is shown in Figure 4.1. Some markers may be missing due to device errors or when a user takes off the device. The diet records contain the type (e.g., lunch) and the amount of nutrients (e.g., sugar) contained for each meal. Some records may be inaccurate or missing. Diabetic individuals were excluded because their metabolism differs extensively from healthy individuals, and the comprehensive modeling of that falls beyond the scope of this work.

Figure 4.1: Visualization of 3-day data for one patient. Red dots represent glucose markers in the training set, while brown ones in the testing set. Diet records are displayed by vertical bars, whose nutrients and their amounts are indicated by different colors. The green line demonstrates the final fitted trajectories, which is a combination of the dashed blue line, a counterfactual trend, and the mean of red lines, samples of estimated treatment response.

4.2 Metrics

The models are trained using data from the first two days, and the third day is used for testing. The performance of treatment-response estimation is quantified using five metrics $\mathrm{M}_i, i \in \{1, \ldots, 5\}$. M_1 is the proportion of variance explained by the trend:

$$\mathrm{M}_1 = \frac{1}{N} \sum_n \frac{\mathrm{Var}(\mathcal{T}_n)}{\mathrm{Var}(y_n)}.$$

M_2 indicates how much more is explained when also the treatment responses are included:

$$\mathrm{M}_2 = \frac{1}{N} \sum_n \frac{\mathrm{Var}(\mathcal{T}_n + \sum_m \mathcal{R}_{nm})}{\mathrm{Var}(y_n)} - \mathrm{M}_1.$$

In detail, a large M_1 means that the outcome is mostly explained by the trend, and a small M_2 represents an inactive treatment response. These metrics are computed in regions of non-zero treatment response. Metrics M_3 and M_4 are simply the mean squared errors in the training and test data. They are calculated for all individuals for whom M_2 indicates that the response has been properly learned. Thus one patient, shown in Figure 4.2, with $\mathrm{M}_2 \approx 0.05$ for the baseline model $\mathcal{M}_{hier}$ is excluded from MSE calculations (other patients have $\mathrm{M}_2 > 0.3$).

Because M_4 measures point-by-point error, it may give misleadingly low values even if the shape of a response is correct if its location is inaccurate. Therefore, M_5 is included and it is insensitive to the inaccuracy in location, and it measures the absolute error in variance between predicted response and outcome:

$$\mathrm{M}_5 = \frac{1}{N} \sum_n |\mathrm{Var}\Big(\sum_m \mathcal{R}_{nm}\Big) - \mathrm{Var}(y_n)|$$

Because our interest is in estimation of the treatment response, and not in the trend, we calculate M_4 and M_5 in windows including one hour before and three hours after each meal.

We use the Mann-Whitney U-test [Mann and Whitney, 1947] to test if other models are better than $\mathcal{M}_{hier}$ in terms of test error M_4. The reason for using $\mathcal{M}_{hier}$ as the baseline is the main argument of this article that EIV modeling is beneficial when estimating treatment-response trajectories, and $\mathcal{M}_{hier}$ is otherwise the same as $\mathcal{M}_{hier+time}$ and $\mathcal{M}_{hier+time+cov}$ except that it the does not include the EIV components. We also compare the models using an information criterion for predictive accuracy. The state-of-the-art criterion is leave-one-out cross-validation (LOO) [Vehtari et al., 2017], which is used here.

4.3 Results

	M_1 PVE Trend	M_2 PVE Resp	M_3 MSE Train	M_4 MSE Test	M_5 ΔVar Test	p-value U-test	LOO	pLOO	SE LOO
$\mathcal{M}_{ind}$	0.361	0.342	0.149	1.695	0.927	1.00	3550	247	319
$\mathcal{M}_{hier}$	0.359	0.339	0.159	0.752	0.391	-	3588	215	317
$\mathcal{M}_{hier+time}$	0.350	0.424	**0.098**	**0.738**	0.377	3.24e-4	2870	342	265
$\mathcal{M}_{hier+time+cov}$	**0.345**	**0.424**	0.100	0.742	**0.373**	3.00e-6	2948	420	350

Table 4.1: Comparison of models using the real-world glucose data. The metrics M_1 through M_5 are defined in text, where PVE means *Proportion of Variance Explained*. p-value tests if other models are better than $\mathcal{M}_{hier}$ in terms of M_4. LOO stands for leave-one-out cross-validation, pLOO is the estimated effective number of parameters, and SE-LOO records the standard error in the LOO computations.

Result are shown in Table 4.1. We see that all models outperform the non-hierarchical baseline $\mathcal{M}_{ind}$ by a large margin. Furthermore, taking treatment time inaccuracy into account in $\mathcal{M}_{hier+time}$ improves significantly over the non-EIV model $\mathcal{M}_{hier}$. In fact, estimation of the response fails completely for some individuals without time EIV; the results with and without time uncertainty modeling for one such case are shown in Figure 4.2. On the other hand, taking uncertainty in covariates into account does not notably improve accuracy, owing to the increased flexibility and limited amount of data. Overall, models with EIV component outperform the model without EIV in all metrics.

Interpretability of personalized treatment response is also of great interest; for instance, understanding how an individual's glucose level changes if

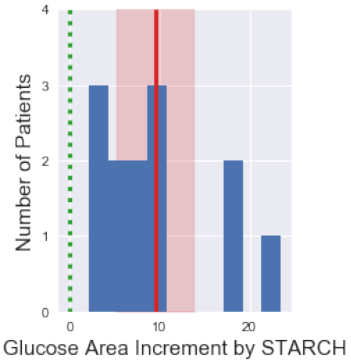
Figure 4.2: Demonstration of time uncertainty modeling for one individual. *Upper:* Results using $\mathcal{M}_{hier+time}$, where arrows indicate the estimated meal times; *Bottom:* Results using $\mathcal{M}_{hier}$.

she eats one more unit of sugar. The overall goal of glucose monitoring is to keep the glucose level in a given range, and both the amount of excess as well as time of staying in hyperglycemic state are clinically important. Hence, a sensible parameter to consider is the impact of different nutrients on the *area* of the response curve. Though this is not a parameter of our model, it is straightforward to derive the personalized increase in response area due to one unit increase of a specific nutrient ΔA_{np} ($n \in 1, \ldots, N$, $p \in \{1...P\}$), using coefficients for height and width, which are modeled explicitly (see Subsection 3.5).

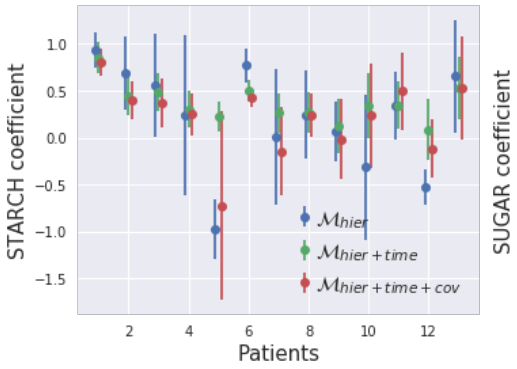
Overall, starch and sugar have the strongest positive impact on glucose (Figure 4.3a), consistent with the understanding that carbohydrates increase blood glucose [Wolever and Miller, 1995]. Protein, on the other hand, has a negative impact, which has been observed before and might represent a complex short-term interaction between nutrients [Karamanlis et al., 2007]. An advantage of our model is that we get *personalized* coefficients for each individual, as shown for starch in Figure 4.3b. Finally, posterior uncertainty of personalized starch coefficients is shown in Figure 4.3c. Importantly, models with EIV have much narrower confidence intervals, meaning that they are estimated more accurately, thanks to increased flexibility that allows fitting the complex data.

	Mean	Std
STARCH	9.57	4.29
SUGAR	6.36	4.97
FIBC	2.83	4.95
FAT	5.29	5.49
PROT	−10.13	3.90

(a)



(b)



(c)

Figure 4.3: *a*). Average impact on response area ΔA_{np} by different nutrients; *b*) Histogram of personalized starch coefficients and their mean (+/- one SD) (red); *c*) Posterior uncertainty in the personalized starch coefficients.

Chapter 5

Discussion

While our model demonstrates superior performance in the task of estimating personalized treatment-response trajectories, there are many future directions to improve it further. First, a more fine-grained measurement error model for treatment covariates can be exploited to integrate EIV with domain knowledge, thus improving identifiability. Second, G-formula can be utilized to estimate causal treatment response over an entire sequence of treatments, instead of the most recent one. Third, currently MCMC sampling is applied for the model inference, which is notoriously slow, whereas variational inference is able to scale efficiently and benefits from the increasing amount of data. Fourth, a multiple linear regression on treatment covariates is used to reconstruct the form of treatment-response trajectories, while further taking into account interaction between covariates would contribute a more accurate model and reveal interesting combined treatment effects.

Chapter 6

Conclusions

In this thesis, we propose a novel model to tackle the difficult problem of estimating treatment-response trajectories. Our model takes into account of error in both timing and covariates of treatments, and shares information across patients under a hierarchical architecture in response to data sparseness, and bestows a causal interpretation on the result. Our model is applied to a real-life dataset where patients' blood glucose trajectories are modeled as a combination of a nonparametric trend and a parametric treatment-response function, which shows that hierarchical structure and errors-in-variables improve the predictive accuracy significantly. Moreover, the model result enjoys easy and meaningful interpretability, which would be greatly beneficial to practitioners in reality.

Bibliography

Roy Adams, Yuelong Ji, Xiaobin Wang, and Suchi Saria. Learning models from data with measurement error: Tackling underreporting. *arXiv preprint arXiv:1901.09060*, 2019.

Christopher M Bishop and Tom M Mitchell. Pattern recognition and machine learning. 2014.

Jennie E Brand and Yu Xie. 11. identification and estimation of causal effects with time-varying treatments and time-varying outcomes. *Sociological Methodology*, 37(1):393–434, 2007.

Raymond J Carroll, David Ruppert, Ciprian M Crainiceanu, and Leonard A Stefanski. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

Xiaohong Chen, Han Hong, and Denis Nekipelov. Measurement error models. *Prepared for the Journal of Economic Literature. www. stanford. edu/˜ doubleh/eco273B/surveyjan27chenhandenis-07. pdf*, 2007.

Issa J Dahabreh, Radley C Sheldrick, Jessica K Paulus, Mei Chung, Vasileia Varvarigou, Haseeb Jafri, Jeremy A Rassen, Thomas A Trikalinos, and Georgios D Kitsios. Do observational studies using propensity score methods agree with randomized trials? a systematic comparison of studies on acute coronary syndromes. *European heart journal*, 33(15):1893–1901, 2012.

Robert C Geary. Inherent relations between random variables. In *Proceedings of the Royal Irish Academy. Section A: Mathematical and Physical Sciences*, volume 47, pages 63–76. JSTOR, 1941.

Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, and Rajesh Ranganath. Opportunities in machine learning for healthcare. *CoRR*, abs/1806.00388, 2018. URL `http://arxiv.org/abs/1806.00388`.

Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. 2018.

Paul Gustafson. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. CRC Press, New York, 1 edition, 2004. ISBN 1-58488-335-9.

Matthew D Hoffman and Andrew Gelman. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Jiunn T Hwang. Multiplicative errors-in-variables models with applications to recent data released by the us department of energy. *Journal of the American Statistical Association*, 81(395):680–688, 1986.

Angela Karamanlis, Reawika Chaikomin, Selena Doran, Max Bellon, F Dylan Bartholomeusz, Judith M Wishart, Karen L Jones, Michael Horowitz, and Christopher K Rayner. Effects of protein on glycemic and incretin responses and gastric emptying after oral glucose in healthy subjects. *The American Journal of Clinical Nutrition*, 86(5):1364–1368, November 2007. doi: 10.1093/ajcn/86.5.1364.

Bryan Lim. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems*, pages 7494–7504, 2018.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

James M. Robins Miguel A. Hernán. Causal inference. preprint on webpage at `https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/s`, 2018.

Manoranjan Pal. Consistent moment estimators of regression coefficients in the presence of errors in variables. *Journal of Econometrics*, 14(3):349–364, 1980.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

Olav Reiersøl. Identifiability of a linear relation between variables which are subject to error. *Econometrica: Journal of the Econometric Society*, pages 375–389, 1950.

John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

Susanne M Schennach. Measurement error in nonlinear models: A review. Technical report, cemmap working paper, 2012.

Susanne M Schennach, Yingyao Hu, and Arthur Lewbel. Nonparametric identification of the classical errors-in-variables model without side information. Technical report, cemmap working paper, 2007.

Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pages 1697–1708, 2017.

Hossein Soleimani, Adarsh Subbaswamy, and Suchi Saria. Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. *arXiv preprint arXiv:1704.02038*, 2017.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017.

T M Wolever and J B Miller. Sugars and blood glucose control. *The American Journal of Clinical Nutrition*, 62(1):212S–221S, July 1995. doi: 10.1093/ajcn/62.1.212s.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

David Zeevi, Tal Korem, Niv Zmora, David Israeli, Daphna Rothschild, Adina Weinberger, Orly Ben-Yacov, Dar Lador, Tali Avnit-Sagi, Maya Lotan-Pompan, et al. Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–1094, 2015.